ED 042 813 TM 000 097

| | |
|---|---|
| AUTHOR | Lord, Frederic M. |
| TITLE | The Self-Scoring Flexilevel Test. |
| INSTITUTION | Educational Testing Service, Princeton, N.J. |
| SPONS AGENCY | Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office. |
| REPORT NO | RB-70-43 |
| PUB DATE | Jul 70 |
| NOTE | 13p. |
| EDRS PRICE | EDRS Price MF-$0.25 HC-$0.75 |
| DESCRIPTORS | *Measurement Techniques, Models, Multiple Choice Tests, *Psychometrics, *Test Construction, *Testing, Testing Problems, *Tests |
| IDENTIFIERS | Flexilevel Test |

ABSTRACT

Certain modifications of a conventional test are proposed which force the item difficulty level to adjust automatically to the ability level of the examinee. The modified test is called a flexilevel test. Although different examinees take different sets of items, the scoring method provides comparable scores for all. Furthermore, the test is self-scoring. These advantages are obtained without some of the usual disadvantages of tailored testing. Preliminary results indicate that flexilevel testing is more effective than conventional testing whenever the need to obtain accurate measurement over a wide range of ability would otherwise require an unusually wide spread of item difficulty in a conventional test. (Author/AE)

RB-70-43

# THE SELF-SCORING FLEXILEVEL TEST

Frederic M. Lord

'Office of Naval Research Contract N00014-69-C-0017
Project Designation NR 150-303
Frederic M. Lord, Principal Investigator

Educational Testing Service
Princeton, New Jersey

July 1970

RB-70-43

# THE SELF-SCORING FLEXILEVEL TEST

Frederic M. Lord

Office of Naval Research Contract N00014-69-C-0017
Project Designation NR 150-303
Frederic M. Lord, Principal Investigator

Educational Testing Service
Princeton, New Jersey
July 1970

# THE SELF-SCORING FLEXILEVEL TEST

## Abstract

Certain modifications of a conventional test can force the item difficulty level to adjust automatically to the ability level of the examinee. Although different examinees take different sets of items, the scoring method provides comparable scores for all. Furthermore, the test is self-scoring. These advantages are obtained without some of the usual disadvantages of tailored testing.

# THE SELF-SCORING FLEXILEVEL TEST*

It is well known that for accurate measurement, the difficulty level of a psychological test should be appropriate to the ability level of the examinee. With conventional tests, this goal is achieved only if the examinees are fairly homogeneous in ability. The Scholastic Aptitude Test of the College Entrance Examination Board, for example, could doubtless provide more reliable measurement at particular ability levels if it did not need to cover such a wide range of examinee talent.

Furthermore, in many situations it is psychologically desirable that the test difficulty be matched to the examinee's ability. A test that is excessively difficult for a particular examinee may have a demoralizing or otherwise undesirable effect.

## Tailored Testing

There has recently been increasing interest in "branched", "computer assisted", "individualized", "programmed", "sequential", or tailored testing (see Linn, Rock, & Cleary, 1969; Lord, 1969, 1971, and in press; Wood, 1969). When carefully designed, such testing comes close to matching the difficulty of the items administered to the ability level of the examinee. In order to achieve this, good testing procedure in practical applications may require the following:

1. Development of a large number of items for pretesting, perhaps on the order of several thousand.

*This work was supported in part by contract N00014-69-C-0017, project designation NR 150-303, between the Personnel and Training Research Programs Office, Psychological Sciences Division, Office of Naval Research and Educational Testing Service. Reproduction in whole or in part is permitted for any purpose of the United States Government.

2. A very large pretesting to obtain adequate data for statistical analysis of each item.
3. A possibly dubious but very complex statistical analysis of pretest item data to estimate the necessary item parameters in advance of the main testing.
4. A final pool of 500-5000 selected items, for actual administration.
5. Computer simulations of perhaps a hundred different tailoring strategies and scoring methods, in order to select item-administration and scoring procedures that will provide accurate measurement at all ability levels.
6. Test administration by a computer at terminals equipped with teletypes and visual display devices.
7. Experimental testings and statistical analyses to demonstrate to the testing agency, to skeptical examinees, and to their lawyers that the scoring method is fair, in the sense of assigning approximately the same score to an examinee regardless of which subset of items he happens to take.

Steps 3 and 5 above are based on certain assumptions:

1. The available items all measure just one psychological dimension--there are no subclusters of items.
2. The form of the item characteristic curves is known.
3. Sampling errors and specification errors in the estimates of item parameters are negligible; in particular, they do not bias the scores of individual examinees.

Some simplification in the above requirements for tailored testing could be obtained by two-stage testing--use of a routing test followed by the administration of one of several alternative second-stage tests (Lord, 1969). This would reduce the number of items needed and would eliminate the need for a computer to administer them. However, to obtain comparable scores from different second-stage tests, either expensive equating procedures based on special large-scale administrations or else complicated scoring procedures are required.

## Flexilevel Tests

To a degree, the same result, the matching of item difficulty with ability level, can be achieved without most of the above-mentioned disadvantages. This can be done by making modifications in the directions, in the test booklet, and in the answer sheet of an ordinary conventional test. The modified test will be called a flexilevel test.

Consider a conventional multiple-choice test in which the items are arranged in order of difficulty. The general idea of a flexilevel test is simply that the examinee starts with the middle item in the test and proceeds, taking an easier item each time he gets an item wrong, a harder item each time he gets an item right. He stops when he has answered half the items in the test.

Let us consider a concrete example, starting with a conventional test of $N = 75$ items. For purposes of discussion, we assume that the items are arranged in order of difficulty; however, it will appear later that any rough approximation to this is adequate. The middle item of the conventional test (formerly item 38) is the first item in the flexilevel test. It is printed in the center at the top of the first page of the flexilevel test. The page below this, and subsequent pages, are divided in half vertically (see Fig. 1). Items formerly numbered 39, 40, 41,...,75 appear in that order in the right-hand columns, the hardest item (formerly item 75) at the bottom of the last page. In place of the old numbers, these items are numbered in blue as items 1, 2, 3,...,37, respectively. Items formerly numbered 37, 36, 35,...,1 appear in that order in the left-hand columns, the easiest item (formerly item 1) at the bottom of the last page. In place of the old numbers, these numbers are numbered in red as items 1, 2, 3,...,37, respectively (the easiest item is now at the end and is numbered 37). The layout is indicated in Figure 1.

The answer sheet used for a flexilevel test must inform the examinee whether each answer is right or wrong. When the examinee chooses a wrong answer, a red spot appears where he has marked or punched the answer sheet. When he chooses a right answer, a blue spot appears. Answer sheets similar to this are commercially available in a variety of designs.

In answering the test, the examinee must follow one rule. When his answer to an item is correct, he should turn next to the lowest-numbered "blue" item not previously answered. When his answer is incorrect, he should work next on the lowest-numbered "red" item not previously answered.

Each examinee is to answer just $\frac{1}{2}$ (N + 1) = 38 items. One way to make it apparent to him when he has finished the test would be to print the answer sheet in two columns, using the same format as in Figure 1 but with the second column inverted. Thus, the examinee works down from the top in the first column of the answer sheet and up from the bottom in the second column. The examinee can be told to stop (he has completed the test) when he has responded to one item in each row of the answer sheet.

It is now clear that the high-ability examinee who does well on the first items in the test will automatically be administered a harder set of items than the low-ability examinee who does poorly on the first items. Within limits, the flexilevel test automatically adjusts the difficulty of the items administered to the ability level of the examinee.

## Scoring Flexilevel Tests

Let us first agree that when examinees answer the same items, we will be satisfied to use the usual number-right score and to consider examinees with the same number-right score equal. A surprising feature of the flexilevel test is that even though different examinees take different sets of items, complicated and expensive scoring or equating procedures to put all examinees on the same score scale are not necessary. Most important, the simplicity and obvious validity of the scoring will prevent examinees from feeling that they are the victims of occult scoring methods. Finally the test is self-scoring--the examinee can determine his score without counting up the number of correct answers.

A flexilevel test has the following properties, which the reader should verify for himself.

1. If the items were rearranged in order of difficulty, the items answered by a single examinee would always be a block of consecutive items.

For convenience of exposition, assume that the examinee has had time to complete the required $\frac{1}{2}(N + 1) = 38$ items. Also that he has been instructed to indicate on the answer sheet the item he would have to answer next if the test were continued. (In an exceptional case, this might be a dummy "item 38", which need not actually appear in the test booklet, since no one will ever reach it.) An examinee who indicates a blue (red) item will be called a blue (red) examinee.

2. For a blue examinee, the number of right answers is equal to the serial number of the item that would be answered next if the test were continued.

3. For a red examinee, the number of <u>wrong</u> answers is equal to the serial number of the item that would be answered next if the test were continued. The number of right answers is obtained by subtracting this serial number from $\frac{1}{2}(N + 1)$. (A different serial numbering of the red items could give the number of right answers directly, but might confuse the examinee while taking the test.)

4. All blue examinees who have a given number-right score have answered the same block of items.

5. All red examinees who have a given number-right score have answered the same block of items.

It can now be seen that all blue examinees can properly be compared with each other in terms of their number-right scores, even though examinees with different scores have not taken the same test. Consider two blue examinees, A and B , whose number-right scores differ by 1. The higher-scoring examinee, A , is clearly the better of the two because he took the harder test. The items answered by the two examinees are identical except that A had one item that was harder than any of B's and B had one item that was easier than any of A's. From this, it appears that A could still be considered better than B even if the difficulty levels of individual items were subjectively estimated rather than determined by complex statistical procedures.

The same reasoning shows that all red examinees can properly be compared with each other in terms of their number-right scores.

6. Examinees of the same color are properly compared by their number-right scores.

It remains to be shown how blue examinees can be compared with red examinees.

Consider a red examinee with a number-right score of $x$. If his very last response had been correct instead of wrong, he would have been a blue examinee with a score of $x + 1$. Clearly, his actual performance was worse than this; so we conclude that

7. a blue examinee with a number-right score of $x + 1$ has outperformed all red examinees with scores of $x$.

Finally, we can compare a blue examinee and a red examinee, both having the same number-right score. Suppose we hypothetically administer to each examinee the item that he would normally take if the testing were continued. If both examinees answer this item correctly, they both become blue examinees with identical number-right scores. We have agreed that such examinees can be considered equal. In order hypothetically to reach this equality, however, the blue examinee had to answer a hard item correctly, whereas the red examinee only had to answer an easy item correctly.

Clearly, without the hypothetical extra item, the standing of the blue examinee is inferior to the standing of the red examinee:

8. a red examinee has outperformed all blue examinees having the same number-right score.

In view of this last conclusion, let us modify the scoring by adding one-half score point to the number-right score of each red examinee. Thus, once we agree to use number-right score for examinees answering the same block of items, we can say that

9. on a flexilevel test, examinee performance is effectively quantified by number-right score, except that (roughly) one-half score point should be added to the score of each red examinee.

If desired, all scores can be doubled to get rid of fractional scores.

## Conclusion

It is clear from the foregoing that to a considerable extent the flexilevel test matches the difficulty level of the items administered to the ability level of the examinee. This result is not achieved without some complication of the test administration. However, the complications are minor compared with those arising in other forms of tailored testing.

Quantitative theoretical investigations of the measurement gains obtainable by flexilevel testing will be carried out and published elsewhere. Preliminary results indicate that flexilevel testing is more effective than conventional testing whenever the need to obtain accurate measurement over a wide range of ability would otherwise require an unusually wide spread of item difficulty in the conventional test. This same conclusion is reasonably apparent from the nature of the flexilevel test, without need for complicated theoretical investigation.

## References

Linn, R. L., Rock, D. A., & Cleary, T. A. The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 1969, *29*, 129-146.

Lord, F. M. A theoretical study of two-stage testing. Research Bulletin 69-95 and ONR Technical Report, Contract N00014-69-C-0017. Princeton, N. J.: Educational Testing Service, 1969.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance*. New York: Harper and Row, in press. Chapter 7.

Lord, F. M. Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, 1971, in press.

Wood, R. The efficacy of tailored testing. *Educational Research*, 1969, *11*, 219-222.